Econometrics

Use of statistical methods to analyse economic data, and to study economic relationships

- Econometric practice

- Econometric theory

Course Aims

To develop understanding of some of the main methods used in econometric practice

- allowing critical evaluation of empirical evidence presented in the economics literature

- and application in your own empirical research

Will also introduce key concepts used in the study and development of these methods

Types of Data

Cross-section - observations on units at some point in time

Examples: weekly earnings; annual incomes

Time-series - observations on same unit at different points in time

Examples: quarterly GDP; weekly exchange rates

Panel data - repeated observations over time on same units (or time-series data for multiple units)

Examples: manufacturing firms; OECD countries

Data sources

Structure of Course

This term - basic methods for cross-section data

Next term (wks 1-6) - methods for time-series data

Microeconometrics module (HT wks 7-8; TT wks 1-2)

a) methods for panel data

b) further methods for micro data ('corners, kinks and holes')

Econometric Software

This term - Stata - access via Manor road server

Next term - OxMetrics - free for Oxford students

Optional 1 day course on Econometrics with Matlab

Why Econometrics?

Classical statistics - controlled experiments

- quantities controlled by the researcher (no 'feedback' from extraneous factors)

- may be repeated (with quantities 'fixed in repeated experiments')

Controlled experiments are rare in economics and other social sciences

Two exceptions:

- randomised control trials, esp in developing countries

- laboratory experiments

Observational Data

Most of the data used in economics is observational rather than experimental

We observe the weekly wages earned by samples of workers, or the evolution of the pound:dollar exchange rate

But we don't get to change the life histories of individuals and observe the effect this has on their wages, or to manipulate interest rates and observe the effect this has on the exchange rate

Why does this matter?

Example: wages/education/ability

Outline for this term

Wks 1-2 - some key concepts in probability theory - building blocks used in study/development of statistical and econometric methods [BN notes]

Wks 3-6 - apply to study widely used methods for estimating linear relationships from cross-section data, and testing claims about such relationships

Wks 7-8 - generalise some of these ideas, and provide a more rigorous treatment of the approach used to approximate the behaviour of estimators and test statistics in large samples [JD]

Weekly classes - key preparation for exam in June next year

Mock exam - at start of HT, based on MT material

Random Variables

Key concept used to describe outcomes which are uncertain

Example - coin toss

- two possible outcomes: heads or tails

- let $X = 0$ if heads and $X = 1$ if tails

- X is a random variable

More generally, a numerical representation or quantitative function of the possible outcomes

After the event, the observed outcome is the realised value or realisation of the random variable

Discrete random variables take a finite number of possible values

Special case with only 2 possible values - binary random variables (e.g. X above)

Examples in economics:

- is an individual employed or not?

- is a country an oil exporter or not?

The possible outcomes from throwing a dice are described by a discrete random variable which takes 6 possible values

Examples in economics:

- which mode of transport (walking, cycling, car, bus, train, ...) does an individual use to travel between home and work?

- which brand of breakfast cereal did I eat this morning?

Continuous random variables may take infinitely many possible values

Examples: height of graduate students in Oxford; time interval between successive transactions on a financial market

Notice that the range of values taken by a continuous random variable may be bounded - heights and time intervals are non-negative, but may take infinitely many positive values

For random variables which may take a finite but vast number of possible values, it is often useful to treat them 'as if' they are continuous

Example: weekly wages, measured in pence

Probabilities

Used to describe the behaviour of random variables

Example - if we toss a fair coin and define the binary random variable X as above, we have $\mathrm{P}(X = 0) = \mathrm{P}(X = 1) = \frac{1}{2}$

Both probabilities lie between zero and one, and sum to one, as there are only 2 possible outcomes

More generally, for discrete random variables, we denote the probability of outcome x using $\mathrm{P}(X = x)$ [sometimes written $\mathrm{Pr}(X = x)$]

Here we use X to denote the random variable, and x to denote a possible outcome

For continuous random variables, we describe the probability of the outcome lying in particular intervals

Examples: P(height $>$ 6 feet) or P(height $>$ 1.85m)

P(£400.00 $\leqslant$ weekly wage $<$ £450.00)

For a discrete random variable Y which may take one of K possible mutually exclusive outcomes denoted by $y_1, y_2, ..., y_K$, the associated probabilities follow rules or **axioms**:

$\mathrm{P}(Y = y_k) \geqslant 0 \quad \text{for } k = 1, 2, ..., K$

$\mathrm{P}(Y = y_1 \text{ or } Y = y_2 \text{ or } ... \text{ or } Y = y_K) = 1$

$\mathrm{P}(Y = y_k \text{ or } Y = y_j) = \mathrm{P}(Y = y_k) + \mathrm{P}(Y = y_j) \quad \text{for } j \neq k$

The latter two statements involve the combined probability of 2 or more of the possible outcomes

If Y denotes the outcome from throwing a fair dice, and we denote the 6 possible outcomes in the usual way, we then have

$$\mathrm{P}(Y = 1 \text{ or } Y = 2) = \mathrm{P}(Y = 1) + \mathrm{P}(Y = 2) = \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{2}{6} = \tfrac{1}{3}$$

$$\mathrm{P}(Y = 2 \text{ or } Y = 4 \text{ or } Y = 6) = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{3}{6} = \tfrac{1}{2}$$

Denoting the first of these possibilities by '$Y \leqslant 2$' and the second of these possibilities by 'Y is even', and noting that these are **not** mutually exclusive, we can also consider combined probabilities of the form

$$\mathrm{P}(Y \leqslant 2 \text{ and } Y \text{ is even}) = \mathrm{P}(Y = 2) = \tfrac{1}{6}$$

Here it is easy to see that $Y = 2$ is the only outcome which satisfies both conditions

$$\mathrm{P}(Y \leqslant 2 \text{ or } Y \text{ is even}) = \mathrm{P}(Y = 1 \text{ or } Y = 2 \text{ or } Y = 4 \text{ or } Y = 6) = \tfrac{4}{6} = \tfrac{2}{3}$$

For these possibilities, which are not mutually exclusive, we have the relation

$$\mathrm{P}(Y \leqslant 2 \text{ or } Y \text{ is even}) = \mathrm{P}(Y \leqslant 2) + \mathrm{P}(Y \text{ is even}) - \mathrm{P}(Y \leqslant 2 \text{ and } Y \text{ is even})$$

Or, letting A denote the outcome '$Y \leqslant 2$' and B denote the outcome 'Y is even', we have

$$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \text{ and } B)$$

This relation could also be stated using the language of sets, with A denoting $Y \in \{1, 2\}$, B denoting $Y \in \{2, 4, 6\}$, 'or' corresponding to the union of 2 sets, and 'and' corresponding to the intersection of 2 sets

Similar properties apply to probabilities which describe the behaviour of a continuous variable - for example, if the random variable W represents weekly wages (in pounds and pence), and we consider the mutually exclusive possibilities

$$P(£400 \leqslant W < £450 \text{ or } £450 \leqslant W < £500)$$

$$= P(£400 \leqslant W < £450) + P(£450 \leqslant W < £500)$$

which gives the probability of W taking a value between £400.00 and £499.99

The axiomatic approach to probability theory was developed in the 1930s by the Russian mathematician Andrey Kolmogorov

It can be stated in much more general (and abstract) terms than we are using here, using measure theory and set theory

Appendix 1.7 in BN notes provides an introduction and references

This more general treatment will not be required to follow the course this year, but will be useful if you plan to take one of the Advanced Econometrics option papers next year, or to pursue research in econometric theory

Samples and Populations

Recording the outcomes from n tosses of a fair coin would give us a sample of data

Denote the realised values $X_1, X_2, ..., X_n$, or X_i for $i = 1, 2, ..., n$, where each X_i takes the value 0 (heads) or 1 (tails), as before, and n is the sample size

For example, with $n = 10$, we might observe 4 heads ($X_i = 0$) and 6 tails ($X_i = 1$)

We refer to $\frac{4}{10}$ as the sample frequency of the outcome heads, and $\frac{6}{10}$ as the sample frequency of the outcome tails

Let the indicator function $I(X_i = 0) = 1$ if $X_i = 0$, and $I(X_i = 0) = 0$ if $X_i \neq 0$

[A different notation for the same indicator function could be $1_{X_i=0}$, as used in BN notes]

[Since in this example X_i is binary and coded $0/1$, it happens to be the case that $I(X_i = 0) = 1 - X_i$]

The sample frequency then corresponds to the sample average of the indicator function

$$\widehat{f}(0) = \widehat{f}(X_i = 0) = \frac{1}{n} \sum_{i=1}^{n} I(X_i = 0)$$

In the previous example, the indicator $I(X_i = 0)$ takes the value 1 in 4 of the 10 cases, and 0 otherwise, giving the sample frequency as $\frac{4}{10}$

Similarly, letting $I(X_i = 1) = 1$ if $X_i = 1$ and $I(X_i = 1) = 0$ if $X_i \neq 1$ [so that here we happen to have $I(X_i = 1) = X_i$], we have

$$\widehat{f}(1) = \widehat{f}(X_i = 1) = \frac{1}{n} \sum_{i=1}^{n} I(X_i = 0) \qquad [= \frac{6}{10} \text{ in example}]$$

The sample frequency is the **sample analogue** of the probability that the underlying discrete random variable takes a particular value. The claim that $P(X = 0) = P(X = 1) = \frac{1}{2}$ reflects a particular belief about the world, that the coin is fair. Such claims/beliefs about the underlying process generating the data are referred to as claims/beliefs about 'the population'

If the coin is (approximately) fair, our probability model will be a pretty accurate representation of the data generating process (dgp)

In economics, we encounter more complex dgps, and often have to work with probability models or statistical models whose foundations are less solid

One role of econometrics is to provide tools which allow the sample data to be used to test the validity of particular claims or beliefs

[Another role of econometrics is to provide tools which allow us to make predictions, based on the sample data we have seen, about situations we have not seen; in a time-series context, this corresponds to making forecasts]

For small sample sizes, finding that the sample frequencies for heads and tails are not exactly one half does not provide strong evidence against the claim that the coin is a fair one (suppose $n = 9$); we will develop tools to test the claim that the true probabilities are one half, using the sample data

As we consider larger and larger sample sizes, then if the coin is fair, we might suspect that the sample frequencies should get closer and closer to one half; we will also show that this intuition is correct, and provide more precise statements which can be used to test claims or hypothesis in large samples

Sample frequencies or sample proportions can also be defined for the realised values of random variables taking values within particular intervals, which also applies for continuous random variables

Suppose I measure the height of each student here, and define $I(\text{height}_i > 6 \text{ feet}) = 1$ if student i is taller than 6 feet, and $I(\text{height}_i > 6 \text{ feet}) = 0$ otherwise. The sample proportion of students over 6 feet is then

$$\frac{1}{n}\sum_{i=1}^{n} I(\text{height}_i > 6 \text{ feet})$$

in a sample of n students

Here my population of interest could be all graduate students in Oxford, but I find it infeasible to measure all their heights [Caveats?]

Example in economics: we use data on weekly wages for a 1% sample of all UK employees (from the Annual Survey of Hours and Earnings) to draw inferences about weekly wages for the population of UK employees

Distribution Functions

Used to describe the behaviour of random variables (population version) or their realised outcome values (sample version)

Population: For the scalar random variable X, the (cumulative) distribution function is

$$F(x) = \mathrm{P}(X \leqslant x)$$

[May also be written as $F_X(x)$, to emphasise that this is the distribution function of the random variable X]

Example - coin toss: $X = 0$ if heads, $X = 1$ if tails

$$\mathrm{P}(X \leqslant 0 - \varepsilon) = 0 \quad \text{for any } \varepsilon > 0$$

$$\mathrm{P}(X \leqslant 0) = \frac{1}{2}$$

$$\mathrm{P}(X \leqslant 1 - \varepsilon) = \frac{1}{2} \quad \text{for any } 1 > \varepsilon > 0$$

$$\mathrm{P}(X \leqslant 1) = 1$$

$$\mathrm{P}(X \leqslant 1 + \varepsilon) = 1 \quad \text{for any } \varepsilon > 0$$

Plotting these values of $F(x)$ against x illustrates the form of the distribution function for a discrete random variable - a step function

$F(x)$ lies between 0 and 1; as we consider higher values of x, $F(x)$ is monotonically non-decreasing, with jumps up, but not jumps down

Points along the $x-$axis where $F(x)$ jumps up are called 'atoms', corresponding to values where the discrete random variable has (non-zero) probability mass ($X = 0$ and $X = 1$ in our example, each with probability $\frac{1}{2}$, which is the amount by which the distribution function jumps up at these points)

Illustrate form of the distribution function for a continuous random variable which may take values corresponding to any real number

$F(x)$ lies between 0 and 1; as we consider higher values of x, $F(x)$ is monotonically non-decreasing

Note that $F(x) \to 0$ as $x \to -\infty$, since $\mathrm{P}(X \leqslant -\infty) = 0$

And that $F(x) \to 1$ as $x \to \infty$, since $\mathrm{P}(X \leqslant \infty) = 1$

More formally, distribution functions satisfy

i) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

ii) $F(x)$ is monotonically non-decreasing in x

iii) $F(x)$ is right-continuous

Conversely, any function $F(x)$ with these 3 properties is a candidate distribution function for some scalar random variables which takes values corresponding to real numbers

Sample counterpart is the **empirical distribution function**, which gives the proportion of the observations for which the variable is less than or equal to some specified value

For example, if $W_1, W_2, ..., W_n$ denote observations on weekly wages for a sample of size n, we define the indicator function $I(W_i \leqslant w) = 1$ if $W_i \leqslant w$ and $I(W_i \leqslant w) = 0$ otherwise, and the empirical distribution function is

$$\widehat{F}(w) = \frac{1}{n} \sum_{i=1}^{n} I(W_i \leqslant w)$$

NB. We can express the proportion of observations with $W_i > w$ as $1 - \widehat{F}(w)$

Similarly, for probabilities, we have $\mathrm{P}(X > x) = 1 - \mathrm{P}(X \leqslant x) = 1 - F(x)$

In any finite sample, we observe a finite number of realised values (whether the underlying random variable is discrete or continuous), so the empirical distribution function will always take the form of a non-decreasing step function, with jumps up at each of the sample values

If the underlying random variable is continuous, the empirical distribution function will look less and less like a step function as the sample size increases

If the underlying random variable is discrete, the empirical distribution function will still look like a step function as the sample size increases

These properties suggest that the empirical distribution function can provide a good approximation to the distribution function of the underlying (discrete or continuous) random variable in large samples; we will establish some more precise properties of this approximation

Density Functions

Density functions are derived from distribution functions

- less general (may not exist)

- may be easier to interpret and work with analytically

Discrete random variables

Let $F(x^-)$ denote the limit of $F(x)$ as we approach x from below (left limit)

More formally, $F(x-h) \to F(x^-)$ as $h \to 0$ for $h > 0$

Or $F(x^-) = \lim_{h \to 0} F(x-h)$ for $h > 0$

Then the (probability) density function or probability mass function is

$$f(x) = \mathrm{P}(X = x) = F(x) - F(x^-)$$

[May also be written as $f_X(x)$, to emphasise that this is the density function of the random variable X]

If X denotes the outcome from tossing a fair coin, as before, there are only 2 values of x for which we have $f(x) > 0$

These are $x = 0$ and $x = 1$, with $f(0) = f(1) = \frac{1}{2}$

These are the 'atoms', with the probability mass at these points being the amount by which the distribution function jumps up

The collection of x values with $f(x) > 0$ is called the **support** of the distribution

Since we obtain the density function from the distribution function by taking a difference, we can reverse the operation to obtain the distribution function from the density function by taking a cumulative sum:

$$F(x) = \mathrm{P}(X \leqslant x) = \sum_{a \leqslant x} f(a)$$

where the sum is taken over all values of $a \leqslant x$, or equivalently over all values in the support of the distribution which satisfy $a \leqslant x$

Any function $f(x)$ with discrete support which satisfies the conditions

$$f(x) \geqslant 0 \quad \text{and} \quad \sum_x f(x) = 1$$

is a candidate density function for some scalar discrete random variable, and implies a corresponding distribution function $F(x)$

Here the sum is taken over all values of x, or equivalently over all values of x in the support of the distribution

NB. More formal treatments may call $f(x)$ defined here a probability mass function but not a density function, and may refrain from using the notation $f(x)$ to denote the probability mass function; other notations used for the probability mass function $F(x) - F(x^-)$ include $\Delta F(x)$ and $\Delta F_X(x)$

Example - Bernoulli distribution: X is a binary random variable, coded 0/1, with

$$\mathrm{P}(X = 1) = \theta \quad \text{and} \quad \mathrm{P}(X = 0) = 1 - \mathrm{P}(X = 1) = 1 - \theta$$

The density function can be written $f(x) = \theta^x(1-\theta)^{1-x}$ for $x = 0, 1$

The binary random variable describing the outcome of a fair coin toss has a Bernoulli distribution with $\theta = 0.5$, denoted Bernoulli(0.5)

Continuous random variables

Provided the distribution function $F(x)$ can be differentiated with respect to x, we obtain the density function by taking this derivative

$$f(x) = \frac{dF(x)}{dx}$$

Rearranging, we obtain $dF(x) = f(x)dx$

The term $f(x)dx$ tells us how much the distribution changes ($dF(x)$) when we consider a very small change in x (dx)

Can think of this as the first-order term in an approximation, which will be accurate for very small changes in x

Conversely, we can obtain the distribution function from the density function (if it exists) by integrating

$$F(x) = \mathrm{P}(X \leqslant x) = \int_{-\infty}^{x} f(a)da$$

A complication is that there are continuous (but not 'absolutely continuous') distribution functions which are not differentiable, in which case the density function does not exist - we will largely ignore this complication in this course

Example - standard normal distribution - a scalar random variable Z with a standard normal distribution has the density function

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$$

which is commonly denoted by $\phi(.)$ rather than the more general $f(.)$

The standard normal density function depends on z^2, and is thus symmetric about $z = 0$

The factor $1/\sqrt{2\pi}$ is a scaling factor which ensures that the density function integrates to one

The standard normal density function has a bell shape

The corresponding distribution function does not have a closed form analytical expression, but can be approximated numerically very accurately; a common notation is

$$\Phi(z) = \mathrm{P}(Z \leqslant z) = \int_{-\infty}^{z} \phi(a)da$$

Tail probabilities of the standard normal distribution play an important role in hypothesis testing, and in formulating confidence intervals for parameters of interest

For example, we have $\Phi(-1.96) = \mathrm{P}(Z \leqslant -1.96) = 0.025$

And, by symmetry, $\Phi(1.96) = \mathrm{P}(Z \leqslant 1.96) = 0.975$

Hence there is probability 0.95 of obtaining a value between ± 1.96

$Z \sim \mathrm{N}(0,1)$ is commonly used to indicate that the random variable Z has a standard normal or $\mathrm{N}(0,1)$ distribution

[This may also be written as $Z \overset{\mathrm{D}}{=} \mathrm{N}(0,1)$; also, if we have $Z \sim \mathrm{N}(0,1)$ and $Y \sim \mathrm{N}(0,1)$, the statement $Z \overset{\mathrm{D}}{=} Y$ indicates that the 2 random variables Z and Y have the same distribution (though they are not the same random variables)]

[To anticipate, the notation $\mathrm{N}(0,1)$ denotes a normally distributed random variable with a particular expected value (0) and a particular variance (1); the behaviour of normally distributed random variables is fully described by the values of these first two moments of the distribution]

Quantiles

Let X be a continuous random variable with a strictly increasing distribution function $F(x)$, so that the inverse function F^{-1} exists

[Recall that for a function g with $g(x) = y$, the inverse function g^{-1} has the property that $g^{-1}(y) = x$]

Consider the value of x such that $F(x) = \tau$, where τ is between 0 and 1

Denote this value of x by q_τ

Different values of τ correspond to different **quantiles** of the distribution

$F(q_{0.5}) = \mathrm{P}(X \leqslant q_{0.5}) = 0.5$, so that $q_{0.5}$ is the median, or 50^{th} percentile

Similarly $q_{0.1}$ is the first decile or 10^{th} percentile; $q_{0.25}$ is the first quartile or 25^{th} percentile; and so on

To find the quantiles, we use the inverse of the distribution function

Since $F(q_\tau) = \tau$, we have $F^{-1}(\tau) = q_\tau$

The inverse of the distribution function is also called the **quantile function**, sometimes denoted as $Q(\tau) = F^{-1}(\tau)$ (or $Q_X(\tau) = F_X^{-1}(\tau)$)

For distribution functions with flat segments or jumps up, the inverse function is not well-defined

In this case the quantiles may not be unique, i.e. there may be multiple values of x for which $F(x) = \tau$

Transformations of random variables

(a) Distributions of transformed random variables are related

Let X be a random variable with the distribution function $F_X(x) = \mathrm{P}(X \leqslant x)$

Let g be a strictly increasing function, so that $Y = g(X)$ is also a random variable, and the inverse function g^{-1} exists

Then we have

$$F_Y(y) = \mathrm{P}(Y \leqslant y) = \mathrm{P}(g(X) \leqslant y) = \mathrm{P}(X \leqslant g^{-1}(y)) = F_X(g^{-1}(y))$$

For example, if $Y = g(X) = \mu + \sigma X$ for $\sigma > 0$, the inverse function is $X = g^{-1}(Y) = \frac{Y-\mu}{\sigma}$, and we have

$$F_Y(y) = F_X\left(\frac{y-\mu}{\sigma}\right)$$

For a continuous random variable with a strictly increasing distribution function, a special case of this result sets $Y = g(X) = F_X(X)$

Note that since the distribution function describes cumulative probabilities, the random variable $Y = F_X(X)$ takes values between 0 and 1

Now

$$F_Y(y) = \mathrm{P}(Y \leqslant y) = \mathrm{P}(F_X(X) \leqslant y) = \mathrm{P}(X \leqslant F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y$$

for values of y between 0 and 1

This result is known as the **probability integral transformation**

The random variable $Y = F_X(X)$ has a **standard uniform distribution**

The corresponding density is $f_Y(y) = 1$

The quantile function is $F_Y^{-1}(\tau) = \tau$, so that we have $q_\tau = \tau$

Conversely, if Y is a random variable with a standard uniform distribution and F is a strictly increasing distribution function, then the random variable $X = F^{-1}(Y)$ has the distribution function F

This result is used to generate pseudo random numbers with desired distributions on computers

(b) Densities of transformed random variables are related

Let X be a continuous random variable with the density function $f_X(x)$ and support $\mathcal{X}$

Let g be a strictly increasing (or strictly decreasing) and differentiable function

Then the random variable $Y = g(X)$ has support $\mathcal{Y} = g(\mathcal{X})$ and the density function

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{\partial g^{-1}(y)}{\partial y} \right| \quad \text{for values of } y \text{ in } \mathcal{Y}$$

For strictly increasing functions, this can be shown by differentiating our previous result for distributions, using the chain rule

This is known as the **change of variable formula** for density functions

For example, if $Y = g(X) = \mu + \sigma X$ for $\sigma > 0$, the inverse function is $X = g^{-1}(Y) = \frac{Y-\mu}{\sigma}$ with the derivative $\frac{\partial g^{-1}(Y)}{\partial Y} = \frac{1}{\sigma} > 0$, and we have

$$f_Y(y) = \frac{1}{\sigma} f_X\left(\frac{y-\mu}{\sigma}\right)$$

More specifically, if $X \sim \mathrm{N}(0,1)$ and $Y = \mu + \sigma X$ for $\sigma > 0$, then

$$f_Y(y) = \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

where ϕ denotes the density function for a standard normal random variable

We say that $Y \sim \mathrm{N}(\mu, \sigma^2)$

The transformation $Y = \mu + \sigma X$ for $\sigma > 0$ is called a location-scale transformation:

- adding the constant term μ changes the location (expected value) of the random variable (unless $\mu = 0$)

- multiplying by the constant term $\sigma > 0$ changes the scale (variance) of the random variable (unless $\sigma = 1$)

Moments

We often summarise the properties of distributions using their moments

The first moment is called the **expected value** or **expectation**

This is a measure of the central tendency of the distribution of a random variable; other measures of central tendency include the median and the mode

Let X be a random variable with the distribution function $F(x)$

The most general definition we consider is

$$\mathrm{E}(X) = \int x dF(x)$$

where the integral is taken over the support of X

For a **continuous** random variable with the density function $f(x)$, we have $dF(x) = f(x)dx$ and this simplifies to

$$\mathrm{E}(X) = \int x f(x) dx$$

For a **discrete** random variable with density (probability mass) function $f(x)$, we have

$$\mathrm{E}(X) = \sum_x x f(x)$$

Expected values are sometimes denoted by $\mathrm{E}_X(X)$, to emphasise that the expectation (sum or integral) is taken over the distribution of the random variable X

These expected values are only defined if the sum or integral is finite; otherwise we say that the expected value does not exist

Example - let $X \sim \text{Bernoulli}(\theta)$

$$\text{E}(X) = (0 \times f(0)) + (1 \times f(1)) = (0 \times (1-\theta)) + (1 \times \theta) = \theta$$

Example - let $X \sim \text{N}(0,1)$. Recall that the standard normal density function is symmetric about zero, so that $\phi(x) = \phi(-x)$. Hence

$$\begin{aligned}\text{E}(X) &= \int_{-\infty}^{\infty} x\phi(x)dx = \int_{-\infty}^{0} x\phi(x)dx + \int_{0}^{\infty} x\phi(x)dx \\ &= -\int_{0}^{\infty} x\phi(x)dx + \int_{0}^{\infty} x\phi(x)dx = 0\end{aligned}$$

NB. The argument used here assumes that this expected value exists

An example of a distribution for which the expected value does not exist is the t-distribution with 1 degree of freedom (Cauchy distribution)

A useful property of expected values is that **the expectation operator is a linear operator**, meaning that we have

$$\mathrm{E}(\mu + \sigma X) = \mu + \sigma \mathrm{E}(X)$$

for any two constants μ and σ (NB. this property is not restricted to $\sigma > 0$)

Example - suppose $X \sim \mathrm{N}(0, 1)$ and $Y = \mu + \sigma X$. Then

$$\mathrm{E}(Y) = \mathrm{E}(\mu + \sigma X) = \mu + \sigma \mathrm{E}(X) = \mu + (\sigma \times 0) = \mu$$

μ is thus the expected value of the random variable $Y \sim \mathrm{N}(\mu, \sigma^2)$

The sample analogue of the expected value is the **sample mean** or sample average

Letting $X_1, X_2, ..., X_n$ denote n observations on a variable X, the sample mean is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

We used the sample mean previously when defining sample frequencies

We will see that the sample mean has some desirable properties as an estimator of the expected value of a random variable

If X is a random variable and g is a function such that $g(X)$ is also a random variable, the expected value of $g(X)$ is

$$\begin{aligned} \mathrm{E}[g(X)] &= \int g(x)dF(x) \\ &= \int g(x)f(x)dx \quad \text{for continuous } X \\ &= \sum_x g(x)f(x) \qquad \text{for discrete } X \end{aligned}$$

Notice that the sums and integrals here are taken over the support of X, not over the support of $g(X)$; this can be made explicit using $\mathrm{E}_X[g(X)]$

This allows us to write the second moment $\mathrm{E}(X^2)$ and higher moments $\mathrm{E}(X^r)$ for $r = 3, 4, ...$; these are sometimes called moments about zero or moments about the origin

The **variance** is obtained by setting $g(X) = [X - \mathrm{E}(X)]^2$, or subtracting the expected value from the random variable itself and then squaring the result; this is sometimes called taking the quadratic variation around the mean (where 'mean' is used in its population sense)

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\{[X - \mathrm{E}(X)]^2\} \\
&= \int [x - \mathrm{E}(X)]^2 f(x) dx \quad \text{for continuous } X \\
&= \sum_x [x - \mathrm{E}(X)]^2 f(x) \qquad \text{for discrete } X
\end{aligned}$$

The variance is the second moment about zero of the random variable $X - \mathrm{E}(X)$; this is also called the second moment about the mean of the random variable X, or the second centred moment of X

The variance gives a measure of dispersion of the random variable around its expected value; other measures of dispersion can be based on quantiles, such as the interquartile range ($q_{0.75} - q_{0.25}$)

Note that **variances are non-negative**, since $[X - \mathrm{E}(X)]^2 \geqslant 0$ for all possible values taken by the random variable X

Indeed, if X is truly a random variable (i.e. not guaranteed to take some constant value), its variance is strictly positive

The variance can also be expressed as

$$\mathrm{Var}(X) = \mathrm{E}\{[X - \mathrm{E}(X)]^2\} = \mathrm{E}(X^2) - [E(X)]^2$$

where the first term is the second moment about zero and the second term

is the square of the expected value

Example - let $X \sim \text{Bernoulli}(\theta)$, and recall that we have $E(X) = \theta$

$$\begin{aligned}
\text{Var}(X) &= (0-\theta)^2 \times f(0) + (1-\theta)^2 \times f(1) \\
&= \theta^2(1-\theta) + (1-\theta)^2\theta \\
&= (1-\theta)[\theta^2 + (1-\theta)\theta] \\
&= (1-\theta)[\theta^2 + \theta - \theta^2] = \theta(1-\theta)
\end{aligned}$$

Also since $0^2 = 0$ and $1^2 = 1$, we have $X^2 = X$ so that $E(X^2) = E(X) = \theta$

and

$$\text{Var}(X) = \theta - \theta^2 = \theta(1-\theta)$$

The variance operator is the expectation of a quadratic function, and hence not a linear operator; instead we have the relation

$$\mathrm{Var}(\mu + \sigma X) = \sigma^2 \mathrm{Var}(X)$$

for any 2 constants μ and σ (NB. this property is not restricted to $\sigma > 0$)

Example - suppose $X \sim \mathrm{N}(0, 1)$ and $Y = \mu + \sigma X$. The variance of a random variable with a standard normal distribution is 1. Then

$$\mathrm{Var}(Y) = \mathrm{Var}(\mu + \sigma X) = \sigma^2 \mathrm{Var}(X) = \sigma^2 \times 1 = \sigma^2$$

σ^2 is thus the variance of the random variable $Y \sim \mathrm{N}(\mu, \sigma^2)$

The square root of the variance is called the **standard deviation**

$$\text{sdv}(X) = \sqrt{\text{Var}(X)}$$

The sample analogue of the variance for a sample of observations $X_1, ..., X_n$ may be written as

$$\frac{1}{n}\sum_{i=1}^{n}\{[X_i - \overline{X}]^2\}$$

or as

$$\frac{1}{n-1}\sum_{i=1}^{n}\{[X_i - \overline{X}]^2\}$$

The latter is sometimes called the sample variance, with the former sometimes called the population variance, although both are statistics that we compute from a sample of data

We will explore the properties of these statistics as estimators of the variance of a random variable

Normalised or **standardised central moments** are defined as

$$\mu_r = \mathrm{E}\left\{\left[\frac{X - \mathrm{E}(X)}{\mathrm{sdv}(X)}\right]^r\right\}$$

μ_3 is a measure of **skewness** or asymmetry of the distribution, with symmetric distributions having $\mu_3 = 0$

μ_4 is a measure of **kurtosis**, indicating the relative weights of the probability in the middle of the distribution and the probability in the tails, or how 'peaked' the density function is

Normally distributed random variables have $\mu_3 = 0$ and $\mu_4 = 3$ (NB. for all values of μ and all values of $\sigma > 0$)

$\kappa_3 = \mu_3$ and $\kappa_4 = \mu_4 - 3$ indicate skewness and kurtosis relative to that of the normal distribution; κ_4 is called **excess kurtosis**

Distributions with relatively more mass in their tails than the normal distribution ($\kappa_4 > 0$) are called leptokurtic or heavy-tailed distributions; these are important in describing the behaviour of returns on financial assets

Independence

For 2 random variables X_1 and X_2, a key concept is independence

X_1 and X_2 are said to be independent if and only if, for all x_1, x_2

$$\mathrm{P}(X_1 \leqslant x_1 \text{ and } X_2 \leqslant x_2) = \mathrm{P}(X_1 \leqslant x_1)\mathrm{P}(X_2 \leqslant x_2)$$

For 2 random variables that are independent, knowing the value taken by one tells us nothing at all about the distribution of the other

The notation $X_1 \perp\!\!\!\perp X_2$ is used to indicate that X_1 and X_2 are independent

This definition of independence extends straightforwardly to 3 or more random variables

Independence is not changed by transforming the random variables

Let $X_1, X_2, ..., X_n$ be a set of n independent random variables

Let $g_1, g_2, ..., g_n$ be a set of functions, such that $Y_i = g_i(X_i)$ is also a random variable, for $i = 1, 2, ..., n$

[We rule out functions such as $g(X_i) = c$ for any value taken by X_i, where c is some known constant]

Then $Y_1, Y_2, ..., Y_n$ are also a set of n independent random variables

Recall that if X_1 and X_2 are independent, then knowing the value taken by X_1 tells us nothing about the distribution of X_2

This property tells us that we can't transform X_1 or X_2 in ways such that knowing the value taken by $g_1(X_1)$ would tell us anything about the distribution of $g_2(X_2)$

Two (or more) independent random variables which have the same distribution are said to be **independently and identically distributed**, or iid

Sums of independent random variables

Let X and Y be 2 independent random variables, and let a and b denote 2 known constants

Expected value

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$$

Variance

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$$

Suppose that $Y_1, Y_2, ..., Y_n$ are independently and identically distributed random variables; note that they have the same expected value $\mathrm{E}(Y_i)$ and the same variance $\mathrm{Var}(Y_i)$ for $i = 1, 2, ..., n$.

Now consider the expected value and variance of the sample mean

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \left(\frac{1}{n}\right) Y_1 + \left(\frac{1}{n}\right) Y_2 + ... + \left(\frac{1}{n}\right) Y_n$$

Since $\overline{Y}$ is the sum of independent random variables, we have

$$\begin{aligned} \mathrm{E}(\overline{Y}) &= \left(\frac{1}{n}\right) \mathrm{E}(Y_1) + \left(\frac{1}{n}\right) \mathrm{E}(Y_2) + ... + \left(\frac{1}{n}\right) \mathrm{E}(Y_n) \\ &= \left(\frac{1}{n}\right) [n\mathrm{E}(Y_i)] = \mathrm{E}(Y_i) \end{aligned}$$

We say that the sample mean $\overline{Y}$ is an **unbiased** estimator of the expected value of the random variable

Similarly

$$\begin{aligned}\text{Var}(\overline{Y}) &= \left(\frac{1}{n^2}\right)\text{Var}(Y_1) + \left(\frac{1}{n^2}\right)\text{Var}(Y_2) + ... + \left(\frac{1}{n^2}\right)\text{Var}(Y_n) \\ &= \left(\frac{1}{n^2}\right)[n\text{Var}(Y_i)] = \frac{\text{Var}(Y_i)}{n}\end{aligned}$$

Importantly, the variance of $\overline{Y}$ falls as the number of observations (n) on the random variable increases; the sample mean accumulates information about the central tendency of the distribution as we add more observations, and we become less uncertain about the true value of $\text{E}(Y_i)$

Indeed, we can note that $\text{Var}(\overline{Y}) \to 0$ as the sample size $n \to \infty$; in the limit, we would know the value of $\text{E}(Y_i)$ [although we can only ever calculate sample means and other statistics using finite sample sizes]